Using Semantic Signatures for Social Sensing in Urban Environments

Krzysztof Janowicz

University of California, Santa Barbara, USA

Grant McKenzie

University of Maryland, USA

Yingjie Hu

University of Tennessee, Knoxville, USA

Rui Zhu

University of California, Santa Barbara, USA

Song Gao

University of Wisconsin, Madison, USA

Abstract

The term *social sensing* describes crowd-sourcing techniques and applications that make use of sensors that are closely attached to humans, e.g., as parts of smartphones, and are either directly or indirectly used to provide sensor observations at a high spatial and temporal resolution. In contrast to typical volunteered geographic information (VGI) applications which rely on the conscious and active contribution of information, most social sensing happens on-the-go, i.e., as by-product of human behavior and interaction with technology. Social Sensing has great potential for many applications in urban planning, transportation, crime prevention, health, and so on. In this paper, we focus on a technique called *semantic signatures* to extract and share high-dimensional data about places. Such semantic signatures reveal how people interact with their environment such as the times they visit places of a certain type, e.g., *Winery*, how they communicate about such places, and how these places are distributed throughout space. We will provide an overview of signatures, methods to extract them, and highlight examples for their usage from previous work ranging from location privacy and the extraction of regions, to reverse geocoding.

Key words: Semantic Signatures, Social Sensing, Cyber-Physical Systems

1. Introduction

Several terms have been introduced over the past years to characterize a broader underlying paradigm shift in the ways research is carried out across many domains ranging from the social to the physical sciences. *Big data*, for instance, highlights the increasing availability of massive datasets which enable researchers to answer new questions by giving access to a higher spatial, temporal, and thematic resolution than before, but requires novel techniques, e.g., parallelization, to handle the size of these data. The related concept of *data science*, focuses on techniques to collect, clean, integrate, analyze, and visualize this data deluge. Several variations of these original terms have been introduced more recently to address some criticism related to big data and data science. For instance, *broad data* and *smart data* are both meant to highlight the fact that *size* alone is of less importance than the heterogeneous sources where such data may come from or the meaningful pre-selection and interpretation of the data [1]. Gray's notion of a *fourth paradigm of science* [2] focuses on how the wide availability of data changes the inner workings of scientific workflows, e.g., by the unexpected/opportunistic reuse of existing data. Finally, others have pointed to the increasing need for techniques to support the meaningful integration and synthesis of datasets given their growing volume, variety, and velocity[3].

Given this broader trend, it is worth asking how these new datasets are created and how insights derived from these data can be made more readily available, i.e., without the need to access the full data. Interestingly, many recent breakthroughs in the broader field of data science are the result of *social machines*, i.e., large-scale, socio-technical systems that arise from the interaction of humans and machines [4, 5]. Typical examples for such systems are Wikipedia, CAPTCHA-like systems to improve optical character recognition (OCR), or massive datasets labeled by human users or via their usage. One increasingly important method for collecting observational data of human behavior and interaction with the environment is *social sensing* [6, 7]. It describes crowd-sourcing techniques and applications that make use of sensors that are closely attached to humans, e.g., as parts of smartphones, and are either directly or indirectly used to provide sensor observations at a high spatial and temporal resolution. While user-generated content (UGC), such as volunteered geographic information (VGI) [8], typically relies on conscious and active contributions, social sensing often utilizes data that are created as by-products of human behavior and their interaction with technology. To give a concrete example, VGI includes tasks such as digitizing streets for the OpenStreetMap (OSM) project, while social sensing may utilize the fact that certain streets or neighborhoods are digitized and updated earlier and more frequently than others or that people visit types of places during characteristic hours or in distinctive sequences [9].

Social sensing offers great potential for applications in urban planning, transportation, health, crime prevention, disaster management, and so on. For instance, social sensing has been proposed as a method for crowdsourced earthquake early warning systems [10]. In this work, we focus on a technique called

semantic signatures to extract and share high-dimensional data about types of places and neighborhoods. Semantic signatures are an analogy to spectral signatures that play a crucial role in remote sensing. While these spectral signatures uniquely identify types, e.g., land cover classes, via characteristic reflectance or emittance patterns in the wavelengths (called *bands*) of electromagnetic energy, semantic signatures utilize data traces from human behavior. Just like libraries of spectral signatures that have been used in fields ranging from agriculture to studying the atmosphere of distant planets, semantic signatures can be used in a variety of ways. In fact, we will discuss examples such as reverse geocoding, geo-privacy, co-reference resolutions, and so forth.

To give an intuitive example, semantic signatures rely on the fact that people frequently go to bakeries during the morning hours and are more likely to mention them in the context of baking, coffee, cakes, sandwiches, and so forth, while nightclubs show very distinct temporal patterns and would unlikely be mentioned in a sentence together with baking. From an inferential perspective, this implies that an unknown place visited during Friday night, co-located with other places visited during evening hours, and mentioned in the context of drinks and dancing is very unlikely to be a bakery, but rather a bar or nightclub. Each data collection and analysis method introduced in the following Section 2 to 4 can be seen as a semantic *band*, and any combination of these bands that uniquely identifies a place type becomes a *signature*. For example, bars and nightclubs may be difficult to tell apart by just looking at the hours and days they are visited, but conversations about bars, e.g., in a local business review, are more likely to mention "sports" or "taps", while these terms are less likely to occur in the context of nightclubs. Hence, combining thematic and temporal data can help uniquely identify place types. It is worth mentioning that many of these distinctions are intuitive to humans, but we need probabilistic models to integrate these distinctions into computational models and workflows. Place types themselves are a key component to geographic information retrieval, recommender systems, urban planning, and so forth, as they are a proxy for the affordances [11], i.e., action potentials, of places and neighborhoods.

The remainder of this chapter is organized as follows. From Section 2 to 4, we present an overview of spatial, temporal, and thematic signatures respectively, and discuss the methods that can be used for extracting these signatures. In Section 5, we outline a variety of examples from previous work to demonstrate the values of these signatures by highlighting their usage. Finally, Section 6 summarizes this chapter and discusses future directions.

2. Spatial Signatures

Spatial signatures capture the characteristics of places through their distributions over geographic space, as places of a given type often have a unique pattern in which they appear and co-locate with other places. For example, the distribution of mountains is likely to be different from that of hotels; and the same comparison can be made for other urban points of interest (POI) such as restaurants and schools.

We adopt a set of spatial statistics, and use them to characterize the semantics of place types. We call the collection of these type-wise statistics a *spatial signature* [12]. These signatures have been employed for tasks such as aligning place types across different gazetteers (e.g., GeoNames, Getty Thesaurus of Geographic Names (TGN) and DBpedia Places) and POI datasets (e.g., Foursquare, Factual, and Google Places) to increase the interoperability across different data sources. A variety of spatial statistics can be adopted to extract spatial signatures. In the following, we describe 4 types of statistics using specific examples. A more comprehensive discussion on many other statistics can be found in our previous work [12].

Spatial Point Pattern. As geographic information in most gazetteers and social media are stored in the format of point-features, i.e., without more detailed geometries, we first describe techniques from spatial point pattern analysis to quantify the point distribution of feature types across a study domain. Both local and global point patterns can be extracted. Regarding local point patterns, both intensity-based (e.g., local intensity and kernel density estimations of local areas) and distance-based analysis (e.g., nearest neighbor analysis, Ripley's K, and standard deviational ellipse analysis) are employed. These statistics are supposed to capture spatial arrangements of points in a local scope. With respect to global point patterns, we compute the points intensity and estimate their kernel density on a global scale to capture their global spatial distribution. Corresponding statistics, such as the range of Ripley's K and the bandwidth of kernel density estimations are selected from these statistics. Figure 1 illustrates a comparison between the place types of *Park* and *Dam* in terms of their point patterns using Ripley's K. It shows that parks in the DBpedia Places dataset are more clustered compared to dams, as the observed curve (solid black line) of parks deviates more from the theoretical one (dotted red line) which is built under complete spatial randomness (CSR).

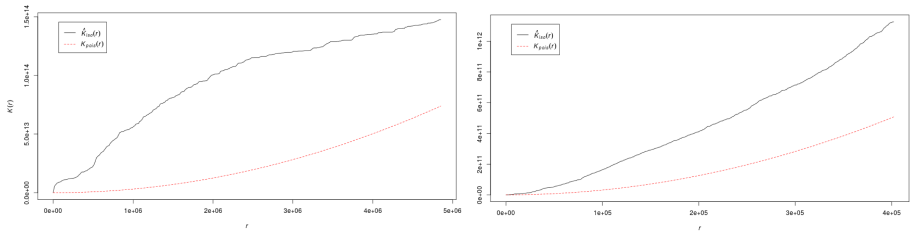


Figure 1: Ripley's K of *Park* (left) and *Dam* (right) from DBpedia Places.

Spatial Autocorrelations. In addition to spatial point pattern analysis in which the distribution of points is the main focus, spatial autocorrelation analysis is adopted with a focus on investigating spatial interactions among features represented by point geometries. Second-ordered interaction analysis, Moran's I, and semivariances, are utilized in this group. Moran's I quantifies how intensities of cells differ from their neighbors, and semivariances measure the variation

of cell intensities in a specific distance lag class. For semivariances, we select values at the first, median, and last distance lags as bands for our spatial signatures as they represent variation on small, median and large scales, respectively. Figure 2 shows that the patterns of spatial autocorrelations (e.g., the nugget,

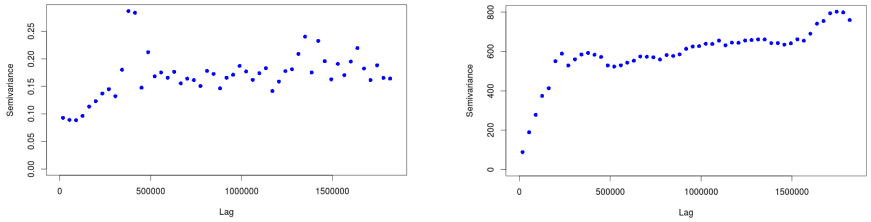


Figure 2: Experimental semivariogram of *Park* (left) and *Dam* (right) from TGN.

range, sill, and the trend) are different between *Park* and *Dam* in TGN.

Spatial Interactions with Other Geographic features. This group of statistics extends spatial signatures to consider the interactions between target place types and other geographic information. Such external information can be population-based, climate-based, or utilizing road networks. One of the reasons to choose these types of data is that they are semantically rich. For instances, features such as mountains are less likely to occur in densely populated areas while the opposite is true for hospitals. Likewise, the frequency distributions of nearest road types for *Amusement Park* and *Restaurant* are significantly different (see Figure 3). Amusement parks are more likely to be located on *avenues*, while restaurants have a higher chances to be located on *roads*.

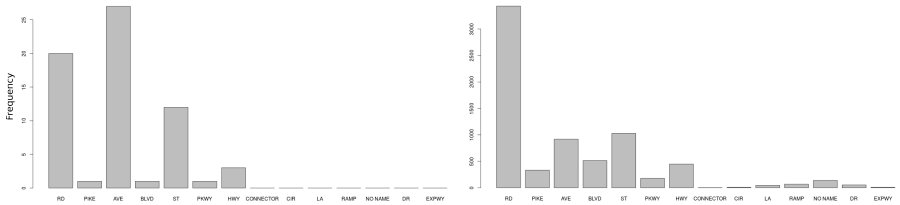


Figure 3: Histogram of road types for *Amusement Park* (left) and *Restaurant* from Google.

Place-based Statistics. In addition to the aforementioned traditional spatial analysis, place-based statistics can be used to characterize the semantics of place types as well. In contrast to spatial statistics, they focus more on describing the topological and hierarchical relations between places. In our case, for example, the number (and entropy) of distinct states (or counties) a place type occurs in, as well as the number (and entropy) of adjacent states (or counties) that also contain features of the target type, are included to indicate the topological relation (e.g., contains and meets) between places and administrative regions.

Table 1: A summary of the 41 statistics for spatial signature.

<table>
<tr><th colspan="2">Spatial Point Pattern</th><th>Spatial Autocorrelations</th><th colspan="2">Spatial Interaction
with Other Geographic Features</th><th colspan="2">Place-based statistics</th></tr>
<tr><td rowspan="11">Local</td><td>Intensity</td><td rowspan="7">Global Moran's I</td><td rowspan="4">Population</td><td>min</td><td colspan="2" rowspan="2">Number of distinct states (or counties)</td></tr>
<tr><td>Mean distance to nearest neighbor</td><td>max</td></tr>
<tr><td>std. of distance to nearest neighbor</td><td>mean</td><td colspan="2">Entropy of states (or counties)</td></tr>
<tr><td>Kernel density (range)</td><td>std.</td><td colspan="2" rowspan="3">Number of adjacent states (or counties) that have the same feature type</td></tr>
<tr><td>Kernel density (bandwidth)</td><td rowspan="5">Road Network</td><td>min of shortest distance</td></tr>
<tr><td>Ripley's K (range)</td><td>max of shortest distance</td></tr>
<tr><td>Ripley's K (mean deviation)</td><td>mean of shortest distance</td><td colspan="2" rowspan="2">Number of distinct feature types for nearest neighbor</td></tr>
<tr><td>std. ellipse (rotation)</td><td rowspan="3">Semivariogram (first distance lag)</td><td>std. of shortest distance</td></tr>
<tr><td>std. ellipse (std. along x-axis)</td><td>entropy of nearest road types</td><td colspan="2" rowspan="2">Entropy of feature types for nearest neighbor</td></tr>
<tr><td>std. ellipse (std. along y-axis)</td><td rowspan="7">Climate</td><td>mean precipitation</td></tr>
<tr><td rowspan="4">Global</td><td rowspan="3">Intensity</td><td rowspan="3">Semivariogram (median distance lag)</td><td>std. precipitation</td><td rowspan="7">LDA-based approach</td><td rowspan="4">Mean KL Divergence of the topic distribtion</td></tr>
<tr><td>mean temperature max</td></tr>
<tr><td>std. temperature max</td></tr>
<tr><td rowspan="2">Kernel density (range)</td><td>mean temperature min</td></tr>
<tr><td rowspan="3">Semivariogram (last distance lag)</td><td>std. temperature min</td><td rowspan="3">Entropy of the topic distribution</td></tr>
<tr><td>Kernel density (bandwidth)</td><td>mean water vapor pressure</td></tr>
<tr><td>std. water vapor pressure</td></tr>
</table>

These statistics are beneficial in terms of distinguishing feature types such as *Glacier* (which occur in eight US-states according to DBpedia) and *River* (which occur in all states). Several other kinds of place-based statistics can be used to uniquely tell apart places of certain types. The used statistics are listed in Table 1.

In summary, spatial signatures are formed by bands extracted from spatial and place-based statistics to uniquely identify place types based on their interactions with other features and alternative sources of geographic data, e.g., climate classifications. Put differently, given a set of statistics about places we can successfully identify their types and we can compare these types, e.g., to study their similarity.

3. Temporal Signatures

Though geospatial properties of place play a key role, there are additional dimensions, or attributes, that help to differentiate places from one another. In fact, it is the combination of properties and attributes that contribute to one's understanding of place. A dimension of place that is of substantial importance to this cause is that of time. There is a temporal component to our definition of place types, one that is reflected in our representation of semantic signatures.

A metro station, for instance, is a very different place at 9am on a Monday than at 3am on a Saturday just as the Roman Colosseum serves a very different purpose today than it did nearly 2000 years ago. The same geographic space can change dramatically depending on the time of day you visit it, day of the week, or season of the year.

The ubiquity of sensor-rich mobile devices has given rise to applications that offer opportunities for users to contribute and share sensor information. Many of these applications and platforms use a gamification model to coerce users into contributing information that can be curated, sold, or analyzed to better understand topics ranging from human urban mobility patterns to health and exercise monitoring. Local business and social platforms such as Yelp and Foursquare offer services that allow users of their platform to *check-in* to the digital representations of local POI [13, 14]. In essence, the process of checking in is the social media equivalent of telling your friends (or the public) that you are at a specific place at a certain time. The underlying place gazetteers consist of rich datasets which contain place attributes ranging from photographs and reviews to curated, user-contributed hierarchies of place types.

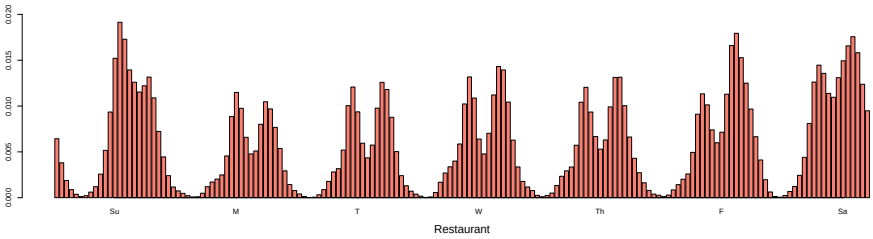


Figure 4: Hourly check-in patterns aggregated to one week for the *Restaurant* place type.

Accessing these check-ins gives urban researchers an unprecedented opportunity to examine the temporal visiting behavior of individuals to a plethora of place types. Through querying data from the public-facing Foursquare application programming interface (API), previous work accessed approximately 3.6 million check-ins to 1 million POI from 421 place types across the United States, United Kingdom, and Australia. Check-in counts per place type were cleaned and aggregated to the nearest hour of day and day of the week. This results in place type specific temporal signatures such as the one shown in Figure 4. This figure demonstrates visiting behavior to restaurants in Los Angeles, CA at an hourly resolution over the course of an average week. We can clearly identify

days of the week based on cyclical daytime vs. night-time patterns (e.g., limited number of check-ins at 2am). The peaks in the figure reflect the typical busy times at a restaurant, namely lunch and dinner time with a slight reduction in the volatility of the popular times on Saturday and Sunday.

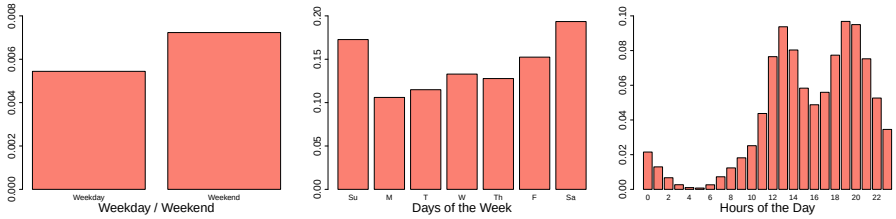


Figure 5: Aggregating temporal signatures at three different scales: (a) Weekdays vs. Weekends (b) Days of the Week (c) Hours of the Day.

These temporal signatures can be further manipulated to explore patterns at a variety of temporal scales. Figure 5 shows the typical *Restaurant* place type visiting behavior for weekdays vs. weekends, days of the week, and hours of the day. Depending on the use case, these temporal aggregates can be used to inform anyone from transit and urban planners to police and commercial entities.

4. Thematic Signatures

So far, we have discussed how models of place can be developed based on their geospatial distributions (i.e., *spatial signatures*) and the temporal characteristics of human-place interactions (i.e., *temporal signatures*). In this section, we take a thematic perspective to formalize place, and will present the concept of *thematic signatures*. In his seminal work [15], Tuan defined *place* as *space* filled with human experience. While human experience is often an intangible concept, people use language to describe their perceptions, feelings, and attachments towards places. Traditionally, many of these human descriptions were in oral form and were ephemeral. In today's big data era, and with the support of various web 2.0 platforms, such descriptions are often automatically recorded in various data sources, such as online reviews (e.g., review comments on restaurants, hotels, and state parks), travel blogs, and social media posts. These large volumes of data enable large-scale, computational studies of human experiences towards places.

Thematic signatures, therefore, aim to capture the characteristics of place types based on the natural language descriptions from people, which serve as a proxy of human experiences. Different places are often situated in different environments and functionalities that afford various sets of human activities [16]. Accordingly, different terms tend to be used by people when describing different places. Intuitively, we are more likely to use terms, such as *hike*, *camping*, *waterfall*, and *nature*, when describing a state park. By contrast, terms such

as *movie*, *popcorn*, *seat*, and *ticket* are more likely to be used when we describe experiences related to cinemas. In relation to *spatial* and *temporal* signatures, *thematic* signatures provide an additional and complementary perspective for understanding and modeling places and their types.

How can we extract such thematic signatures to represent places? The data sources for deriving signatures are descriptive words conveying human experiences. Depending on the way we organize place descriptions, we can extract thematic signatures at both the place-instance level and place-type level. At the place-instance level, we focus on the descriptions for a specific place instance. For example, we can analyze the reviews for a restaurant, *Bob's BBQ joint*, from different people, and learn the main topics that are generally mentioned about this restaurant. At the place-type level, we can aggregate the descriptions for all place instances belonging to the same place type, and extract the thematic signatures for this place type. For example, we can aggregate the reviews for all restaurants in a dataset, and learn the main topics related to the place type *Restaurant*. Aggregated over millions of reviews, these signatures provide a rich representation of place types. Both types of thematic signatures are useful and can be applied to different situations.

A variety of computational models can be utilized to derive thematic signatures for places based on their related natural language descriptions. A simple approach is term frequency and inverse document frequency (TF-IDF) from the field of information retrieval [17]. TF-IDF is based on the bag-of-words model which highlights the words that are used frequently in a document and not very frequently in other documents [18]. For the task of extracting thematic signatures, we can adapt TF-IDF to identify the words that show up frequently in one place instance or place type but not so frequently in other places. The adapted version of TF-IDF is:

$$w_{ij} = tf_{ij} \times \log \frac{|P|}{|P_j|} \tag{1}$$

where w_{ij} is the weight of a term j for place i, tf_{ij} is the frequency of term j used in the descriptions of place i, $|P|$ is the total number of places, and $|P_j|$ is the number of places whose descriptions also contain the term j. Once we have computed the weights for different terms related to a place, word clouds can be employed to visualize the top terms related to a place. These terms with distinct weights can then be used as thematic signatures for places. Figure 6 shows the word clouds based on the reviews of two place types: *Asian Restaurant* and *Stadium*. We can tell the general place types of these two examples even without looking at their place type labels.

Latent Dirichlet allocation (LDA) [19] is a more advanced approach which extracts the major topics associated with different place types. Compared with TF-IDF, LDA is more robust to noise contained in the textual descriptions, handles synonyms, and can capture the semantic relatedness between words. LDA is a generative model which considers each textual document as generated from a probabilistic distribution of topics and each topic as characterized by a distribution over words. As an unsupervised model, LDA discovers semantic

Figure 6: Word clouds for two place types: (a) *Asian Restaurant*; (b) *Stadium*.

topics from the texts without requiring labeled data. Accordingly, each place type or instance can be characterized by the probabilities of different semantic topics based on the related textual descriptions [20, 21]. Figure 7 shows the LDA topic distributions of two place instances: *Right Proper Brewing Co.* and *Moon Under Water Pub & Brewery*. Both examples are of the same place type

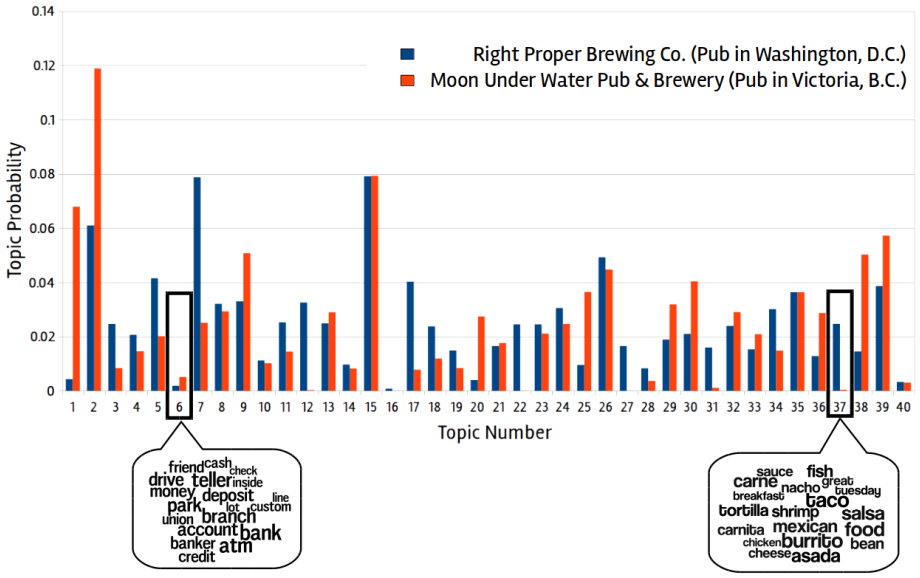


Figure 7: Probability distribution of the LDA topics of two pubs.

(i.e., *Pub*), and thus share similarities in terms of their topics, such as topic 6, topic 13, and topic 24. However, we can also identify the topics under which the two pubs show different characteristics, such as topic 37.

5. Examples

In this section we briefly showcase six examples for studying and applying semantic signatures including basic operations such as place type comparison as well as more specific applications such as improving geo-privacy.

5.1. Comparing Place Types

Due to the semantic heterogeneity of place types, tasks such as query federation, data integration, and conflation become challenging. Therefore, semantic signature has been applied to compare and align place types across different geospatial data sources. In our work, semantic signatures extracted from all three perspectives (i.e., spatial, temporal, and thematic) can be represented as vectors. Let p_1 and p_2 represent two place types, then two vectors can be constructed based on their semantic signatures:

$$p_1 = < f_{11}, \ f_{12}, \ \ldots, \ f_{1D} > \tag{2}$$

$$p_2 = < f_{21}, \ f_{22}, \ \ldots, \ f_{2D} > \tag{3}$$

where f_{ij} represents a (normalized) feature of the semantic signature (e.g., the range of Ripley's K, or the probability of a LDA topic).

With such vector representations, we can measure the semantic similarity between place types using several approaches. One is cosine similarity, which is defined as:

$$s(p_1, p_2) = \frac{\sum_{j=1}^{D} f_{1j} f_{2j}}{\sqrt{\sum_{j=1}^{D} f_{1j}^2} \sqrt{\sum_{j=1}^{D} f_{2j}^2}} \tag{4}$$

Cosine similarity measures the angle of the two vectors constructed from semantic signatures, and is robust to the different magnitudes of values in the vectors. Therefore, cosine similarity is especially suitable for semantic signatures whose vector element values can be largely different. When the vector elements are in probabilities (e.g., topic distribution in thematic signatures), we can also use measurements, such as Jensen-Shannon divergence (JSD), which measures the similarity between two probability distributions. Equation 5 and 6 show the calculation of Jensen-Shannon divergence, where $KLD(P||Q)$ is the KullbackLeibler divergence between two discrete probability distributions P and Q (which are the semantic signatures of the two places to be compared).

$$JSD(P||Q) = \frac{1}{2} KLD(P||M) + \frac{1}{2} KLD(Q||M) \tag{5}$$

$$KLD(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{6}$$

In this section, we demonstrate the usage of both spatial and temporal signatures on comparing place types.

5.1.1. Comparison using Spatial Signatures

Figure 8 depicts a 2D visualization of similarities and differences among place types of three gazetteers: DBpedia Place, GeoNames, and TGN after mapping their high dimensional spatial signatures into 2D using multidimensional scaling (MDS). In general, it can be observed that place types from these gazetteers overlap significantly. Furthermore, three cases are illustrated to show the strength of spatial signatures in aligning specific place types. Case 1 in Figure 8 shows that the parks in DBpeidia Places and TGN are semantically similar compared to the one in GeoNames. This makes sense as the GeoNames gazetteer includes almost all known places such as parks, while DBpedia Places and TGN only record those that are significant in some senses, e.g., historically or culturally. As another example, Case 2 demonstrates that although the same label of a specific place type is shared by the three gazetteers, *Mountain* in this case, their semantics do not align with each other. This case is common across different geospatial ontologies as they are mostly designed by domain experts with certain applications or domains in mind. Semantic signatures have the ability to quantify such *ontological commitments*. In Case 3, three place types that have totally different labels, i.e., *AMD2*, *County*, and *AdministrativeRegion*, are shown to be semantically similar, all representing countries, when using spatial signatures. Such alignments are difficult to establish if only string matching and structural similarities are considered. In summary, by using spatial signatures, one can quantify and subsequently improve the alignment of place types across geospatial ontologies and gazetteers.

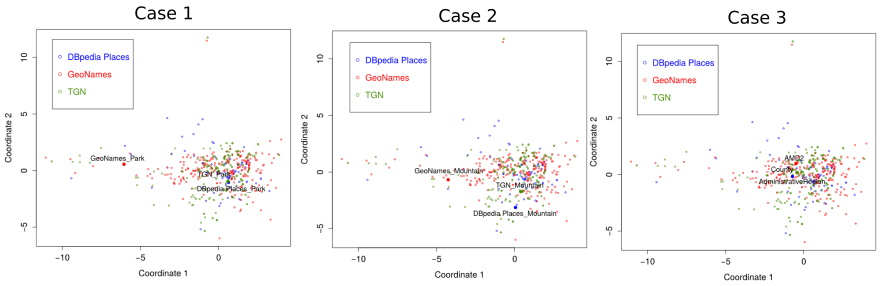


Figure 8: 2D visualization of the alignment of place types across DBpedia Places, GeoNames and TGN. Case 1: *Park*, Case 2: *Mountain*, Case 3: *County*. (Each dot represents a place type)

5.1.2. Comparison using Temporal Signatures

Exploring different place types, we find that many place types have a unique temporal signature and that these signatures can in fact be used to assess the similarity between place types. Figure 9 shows the hourly pattern for airports. Compared to the *Restaurant* temporal signature, airports are relatively *a-temporal* with few peaks throughout the day and limited access at night.

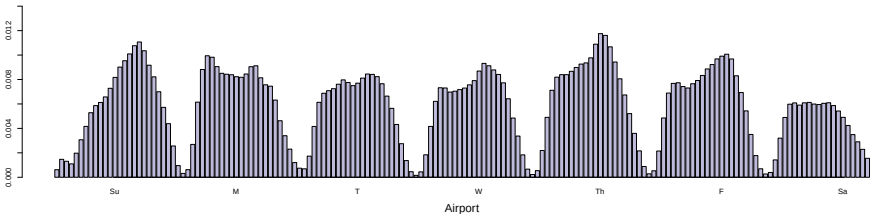


Figure 9: Hourly check-in patterns aggregated to one week for the *Airport* place type.

To give another example, the temporal signature for *Church* shows a clear increase in popularity on Sunday morning with smaller peaks on Sunday and Wednesday evenings.

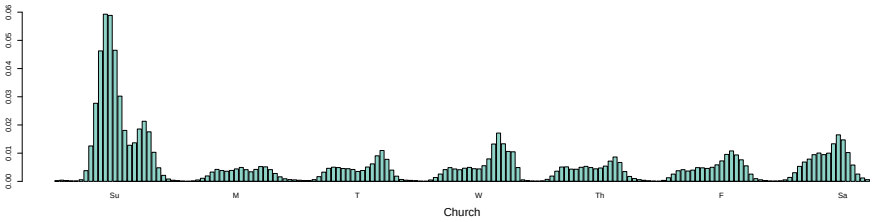


Figure 10: Hourly check-in patterns aggregated to one week for the *Church* place type.

Through assessing the similarity of the temporal signatures, we achieve a better understanding of urban visiting behavior as well as an appreciation of the complexity of modeling the urban landscape. With the goal of better understanding the role that these place types play in defining the city, we developed the *POI Pulse*[1] observatory, a web-based visual platform for exploring interaction between people and places within the city of Los Angeles, CA [22]. In this work, the similarity between the temporal signatures is assessed along with both the geospatial (Section 2) and thematic (Section 4) signatures using information entropy to classify the numerous place types into lower level categories. These lower level categories provided the foundation on which to visually depict the pulse of the city through marker opacity and color.

5.2. Co-reference Resolution across Gazetteers

In addition to aligning place types using spatial signatures, which is discussed in 5.1.1, this subsection outlines an approach for using spatial signatures to match individual geographic features between gazetteers, named as co-reference resolution. In addition to conventional approaches, such as string and structure matching, spatial signatures can be adopted to capture the fact that places have a spatial context that can be used as part of the co-reference resolution

[1] `http://poipulse.com/`

process[23]. The city of *Kobani, Syria* is selected here as an example to illustrate the power of spatial signature, due to its military and geographic importances but also its high ambiguity in different gazetteers (i.e., there are several dissimilar toponyms for Kobani including *Aarab Peunar*, *Kobane*, and *'Ayn al' Arab*). The type-level and instance-level spatial signatures can be combined to match the *Kobani* from DBpedia Places to GeoNames which has in total 5 candidates that cannot be easily matched using conventional approaches. Euclidean distance is used to compute the dissimilarity between candidates and the target in terms of their representations using spatial signature; the candidate that has the smallest dissimilarity to the target is regarded as the match. Experiments show that although one of the candidates in GeoNames is also labeled as *PopulatedPlace*, by taking spatial signatures into account, the *Kobani* in DBpedia Places, labeled as *PopulatedPlace*, can still be correctly matched to *Ayn al 'Arab* in GeoNames, which is labeled as *Seat of a Second-order Administrative Division* [23].

5.3. Geoprivacy

Concerns over location privacy have seen a resurgence in recent years. Mobile devices today are ubiquitous and the sensors available on these devices allow for the collection and distribution of a wide variety of contextual information. In combination with the social web, private information is being shared and distributed at an alarming volume and velocity with arguably little understanding as to the ramifications. The concept of semantic signatures sits very much in the midst of this concern over the sharing of private data as much of the digital footprints that we leave can be curated and compared to the platial data signatures that have been extracted from millions of online sources. For instance, it is possible to substantially limit the possible locations that someone may be at purely based on the textual data that they choose to share online. A microblog post containing the text *"looking forward to burritos and tequila"* posted at 5pm on a Friday in Los Angeles, for instance, provides a high amount of information that can be matched against our probabilistic signatures. The text itself contains references to *Mexican* food and alcohol while the timing of the post indicates a likelihood that the person posting the material will be going to a restaurant rather than a nightclub. Accessing the plethora of freely available gazetteers we can limit the possible locations for the person that created the post [24]. Such an approach does not require access to actual geographic location information. Following the same thought process, signatures can also be used to foster geo-privacy, namely by showing which terms and times are most indicative of a certain activity and place. For instance, replacing 'tequila' with 'drinks' and sending out the message an hour before may increase information entropy to a degree where identifying a place may be less likely [24]. Further work in this area has focused on spoofing one's location and interests based on the inclusion of contextually-relevant *noise* [25] while previous work has focused on the obfuscation of personal identifiable information [26].

5.4. Temporally-enhanced Geolocation

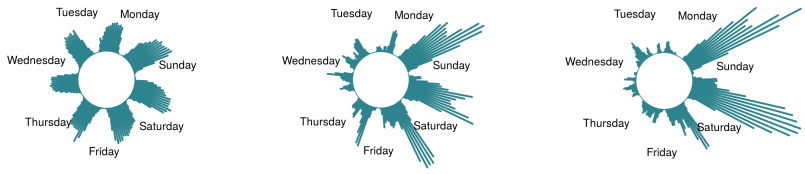
Information related to the temporal dimension of places can be useful in a number of everyday scenarios as well. Take, for example, the process of geolocating, or reverse geocoding. This is a geographic querying method that is executed by millions of people a day as they request the nearest place instances to them based on provided geographic coordinates. Standard approaches to geolocating take a pair of latitude and longitude coordinates (e.g., from a GPS-enabled mobile device) and return a set of nearby places (e.g., Dan's Automotive Shop or Handlebar Coffee Shop). The problem with this approach, however, is that it makes the erroneous assumption that one has the same likelihood of being at a place, regardless of the time of day or day of the week. In actuality, we know that the probability of somebody being at a pub on a Friday at 11pm is significantly higher than the probability of a person being at the Department of Motor Vehicles. Socio-institutional affordances [27] aside, temporal signatures generated from the visiting behavior of millions of individuals clearly demonstrate that there are unique temporal patterns in how people interact with different place types.

Exploiting these temporal patterns, existing work shows that traditional distance-only based approaches to reverse geocoding can be augmented through the inclusion of these time-based probabilistic models [28]. In fact, we show, through a comparison of various methods for including time in such a process, that a temporally-enhanced geolocation method can improve upon the accuracy of the distance-only based method by over 24% (based on a Mean Reciprocal Rank assessment). Such work in combination with others has led to the addition of *Popular Times* plots being included in major mapping and local business platforms [29].

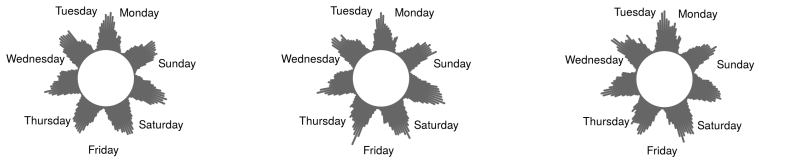
5.5. Regional Variation

The value of temporal signatures built from geosocial visiting behavior in a single city such as Los Angeles, CA is one thing, but building temporal signatures for cities around the world is different in that there will be cultural differences. The question remains as to the uniqueness of place type interactions depending on region. Using check-in data collected from across the United States, Australia and the United Kingdom, the check-ins are split by major cities. Focusing on the cities of Los Angles, CA, Chicago, IL, and New York City, NY, we find that there are significant differences in how the inhabitants interact with place types. Using the Watson's Two-Sample Test, we show that approximately 50% of place types vary significantly ($p < 0.05$) in their temporal signatures [30], while others remain invariant. Figure 11 shows temporal signatures for two place types split by U.S. city. What this work demonstrates is that the temporal visiting behavior of some place types is *a-spatial*, e.g., Drug Stores, while other are regionally variant, e.g., Theme Parks. Additional research on cities outside the United States, namely Sydney, Australia, and London, UK, support these findings, on a more restricted place type dataset. These result also confirm previous research on the habitual behavior of humans in an urban

setting. The findings that roughly 50% of place type temporal behavior is *a-spatial* is of importance for the usefulness of such signatures as well, as it implies that only half of these temporal signatures have to be acquired at a local level for tasks such as reverse geocoding mentioned in Section 5.4 while the other 50% opf place types can be well represented using a single, global signature.



(a) Los Angeles (b) Chicago (c) New York City



(d) Los Angeles (e) Chicago (f) New York City

Figure 11: Circular plots depicting hourly temporal signatures for *Theme Park* (a,b,c) and *Drugstore* (d,e,f).

5.6. Extraction of Urban Functional Regions

Cities support a variety of human activities including living, working, shopping, eating, socializing, and recreation, which usually take place at different types of POI. Compared to other datasets and methods in remote sensing and field mapping, using POI data, social media etc., and associated social sensing methods can lead to a better understanding of individual-level and group-level utilization of urban space at a fine-grained spatial, temporal, and thematic resolution[7]. We use POI that support specific types of human activities on the ground as a proxy to delineate regions with various co-location patterns of POI types [31]. The same type of POI can be located in different land use types and may also support different functions. For example, restaurants are found in residential areas, in commercial areas, as well as in industrial areas. The main function of the POI-type *University* is education, but they also support sports activities, music shows, and so on. We argue that the semantic signatures of POI types can be employed to derive latent classification features, which will then enable the detection and the abstraction of higher-level *functional* regions (i.e., semantically coherent areas of interest) such as shopping areas, business districts, educational areas, and art zones in cities. We collected large-scale datset of Foursquare venues and associated user check-in data in the most populated U.S. cities. Based on the aforementioned data processing procedures and

the LDA topic models by incorporating the popularity score based on unique Foursquare check-in users, we can infer the probabilistic combination of different topics composing a urban function for a region given POI type co-occurrence patterns. For the city of Denver, for instance, we were able to discover [31] a high relevance of the topic *Topic 25*, which consists of a variety of prominent POI types such as *art museum, art gallery, history museum, concert hall and American restaurant*. Such a place may serve multiple functions. The second most prominent LDA topic in this region is *Topic 121* that contains a large percentage composition of *brewery places*. In fact, the region in Denver for which the signatures revealed a dominance of these topics is known as the *"Santa Fe Dr."*, an "Art District" which attracts many local residents, artists, and tourists. This example illustrates the inference capability of our method in identifying urban functional regions given thematic signatures.

6. Summary

In this work we have presented an overview of spatial, temporal, and thematic signatures by discussing the utilized data, the methods to compute and compare signatures, and by providing a variety of examples from our previous work. These examples range from reverse geocoding, neighborhood extraction, co-reference resolution, and ontology alignment to geo-privacy. We have also addressed the question of how local these signatures are, i.e., whether their quality decays when applied to other geographic regions. The results depend on the studied place types and while some types show high variation, others do not. Consequently, global signatures can be augmented with locally trained data to improve results. Recently, there has been increased interest in utilizing embedding techniques to compare place types [32, 33] and early results show that these techniques yield results that strongly correlate with human similarity judgments. Utilizing such techniques for the creation of semantic signatures will be one of the directions for future work.

References

[1] A. Sheth, Transforming big data into smart data: Deriving value via harnessing volume, variety, and velocity using semantic techniques and technologies, in: Data Engineering (ICDE), 2014 IEEE 30th International Conference on, IEEE, 2014.

[2] T. Hey, S. Tansley, K. M. Tolle, et al., The fourth paradigm: data-intensive scientific discovery, Vol. 1, Microsoft research Redmond, WA, 2009.

[3] K. Janowicz, F. Van Harmelen, J. A. Hendler, P. Hitzler, Why the data train needs semantic rails, AI Magazine.

[4] J. Hendler, T. Berners-Lee, From the semantic web to social machines: A research challenge for ai on the world wide web, Artificial Intelligence 174 (2) (2010) 156–161.

[5] N. R. Shadbolt, D. A. Smith, E. Simperl, M. Van Kleek, Y. Yang, W. Hall, Towards a classification framework for social machines, in: Proceedings of the 22nd International Conference on World Wide Web, ACM, 2013, pp. 905–912.

[6] C. C. Aggarwal, T. Abdelzaher, Social sensing, in: Managing and mining sensor data, Springer, 2013, pp. 237–297.

[7] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, L. Shi, Social sensing: A new approach to understanding our socioeconomic environments, Annals of the Association of American Geographers 105 (3) (2015) 512–530.

[8] M. F. Goodchild, Citizens as sensors: the world of volunteered geography, GeoJournal 69 (4) (2007) 211–221.

[9] M. Ye, K. Janowicz, C. Mülligann, W.-C. Lee, What you are is when you are: the temporal dimension of feature types in location-based social networks, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2011, pp. 102–111.

[10] Q. Kong, R. M. Allen, L. Schreier, Y.-W. Kwon, Myshake: A smartphone seismic network for earthquake early warning and beyond, Science advances 2 (2) (2016) e1501055.

[11] T. Jordan, M. Raubal, B. Gartrell, M. Egenhofer, An affordance-based model of place in gis, in: 8th Int. Symposium on Spatial Data Handling, SDH, Vol. 98, 1998, pp. 98–109.

[12] R. Zhu, Y. Hu, K. Janowicz, G. McKenzie, Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics, Transactions in GIS 20 (3) (2016) 333–355.

[13] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, An empirical study of geographic user activity patterns in foursquare., ICwSM 11 (70-573) (2011) 2.

[14] M. Li, R. Westerholt, H. Fan, A. Zipf, Assessing spatiotemporal predictability of lbsn: a case study of three foursquare datasets, GeoInformatica`doi:10.1007/s10707-016-0279-5`.

[15] Y.-F. Tuan, Space and place: The perspective of experience, U of Minnesota Press, 1977.

[16] J. J. Gibson, The theory of affordances, 1977.

[17] C. D. Manning, P. Raghavan, H. Schütze, et al., Introduction to information retrieval, Vol. 1, Cambridge university press Cambridge, 2008.

[18] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, S. Prasad, Extracting and understanding urban areas of interest using geotagged photos, Computers, Environment and Urban Systems 54 (2015) 240–254.

[19] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[20] B. Adams, G. McKenzie, M. Gahegan, Frankenplace: interactive thematic mapping for ad hoc exploratory search, in: Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2015, pp. 12–22.

[21] G. McKenzie, K. Janowicz, The effect of regional variation and resolution on geosocial thematic signatures for points of interest, in: The Annual International Conference on Geographic Information Science, Springer, 2017, pp. 237–256.

[22] G. McKenzie, K. Janowicz, S. Gao, J.-A. Yang, Y. Hu, POI Pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data, Cartographica: The International Journal for Geographic Information and Geovisualization 50 (2) (2015) 71–85.

[23] R. Zhu, K. Janowicz, B. Yan, Y. Hu, Which kobani? a case study on the role of spatial statistics and semantics for coreference resolution across gazetteers, in: International Conference on GIScience Short Paper Proceedings, Vol. 1, 2016, pp. 1–4.

[24] G. McKenzie, K. Janowicz, D. Seidl, Geo-privacy beyond coordinates, in: The 19th AGILE Conference on Geographic Information Science, lecture notes in geoinformation and cartography Edition, Vol. Geospatial Data in a Changing World, Springer, Springer, Helsinki, Finland, 2016, pp. 157–175.

[25] V. Zakhary, C. Sahin, T. Georgiou, A. El Abbadi, Locborg: Hiding social media user location while maintaining online persona, in: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2017, p. 12.

[26] M. Duckham, L. Kulik, A formal model of obfuscation and negotiation for location privacy, in: International conference on pervasive computing, Springer, 2005, pp. 152–170.

[27] M. Raubal, H. J. Miller, S. Bridwell, User-centred time geography for location-based services, Geografiska Annaler: Series B, Human Geography 86 (4) (2004) 245–265.

[28] G. McKenzie, K. Janowicz, Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures, Computers, Environment and Urban Systems 54 (2015) 1–13.

[29] F. Lardinois, Google can now tell you how busy a place is before you arrive, https://techcrunch.com/2016/11/21/google-can-now-tell-you-how-busy-a-place-is-before-you-arrive-in-real-time/, online; accessed 29 January 2018 (2016).

[30] G. McKenzie, K. Janowicz, S. Gao, L. Gong, How where is when? on the regional variability and resolution of geosocial temporal signatures for points of interest, Computers, Environment and Urban Systems 54 (2015) 336–346.

[31] S. Gao, K. Janowicz, H. Couclelis, Extracting urban functional regions from points of interest and human activities on location-based social networks, Transactions in GIS 21 (3) (2017) 446–467.

[32] B. Yan, K. Janowicz, G. Mai, S. Gao, From itdl to place2vec–reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts, Proceedings of SIGSPATIAL 17 (2017) 7–10.

[33] A. Cocos, C. Callison-Burch, The language of place: Semantic value from geospatial context, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Vol. 2, 2017, pp. 99–104.